Transactions in GIS, 2010, 14(1): 43–61

Research Article

Grid-enabling Geographically Weighted Regression: A Case Study of Participation in Higher Education in England

Richard Harris
School of Geographical Sciences and Centre for Market and Public Organisation University of Bristol

Alex Singleton
Department of Geography University College London

Daniel Grose
Centre for e-Science Lancaster University

Chris Brunsdon
Department of Geography, University of Leicester

Paul Longley
Department of Geography University College London

Abstract
Geographically Weighted Regression (GWR) is a method of spatial statistical analysis used to explore geographical differences in the effect of one or more predictor variables upon a response variable. However, as a form of local analysis, it does not scale well to (especially) large data sets because of the repeated processes of fitting and then comparing multiple regression surfaces. A solution is to make use of developing grid infrastructures, such as that provided by the National Grid Service (NGS) in the UK, treating GWR as an "embarrassing parallel" problem and building on existing software platforms to provide a bridge between an open source implementation of GWR (in R) and the grid system. To demonstrate the approach, we apply it to a case study of participation in Higher Education, using GWR to detect spatial variation in social, cultural and demographic indicators of participation.

Address for correspondence: Richard Harris, School of Geographical Sciences University of Bristol, University Road, Bristol, UK BS8 1SS. E-mail: rich.harris@bris.ac.uk

doi: 10.1111/j.1467-9671.2009.01181.x

1 Introduction

Geographically Weighted Regression (GWR) is a method of statistical regression that allows modelled relationships to vary across geographical space (Fotheringham et al. 2002). The technique has been demonstrated successfully in a variety of applications, including health (Nakaya et al. 2005), housing (Bitter et al. 2007), education (Fotheringham et al. 2001), regional economics (Huang and Leung 2002) and environmental applications (Brunsdon et al. 2001).

A constraint on the use of GWR is that its computational requirements grow exponentially with the length of the data set. This is an example of the "big n" problem described by Finley et al. (2009, p. 2874) where ". . . fitting involves matrix decompositions whose complexity increases as $O(n^3)$ in the number of locations, n." This problem can create a trade-off between the size (the geographical extent) of the study and the level of detail (the data resolution) required. For example, high resolution micro-data can be aggregated into fewer and coarser geographical units that are faster to process but lose geographical detail. Alternatively, the data can be partitioned into smaller sub-regions and analysed separately but may induce boundary and other zoning effects by doing so. A third and more desirable option is to overcome the computational problems through parallel implementation of the GWR software using grid or distributed computing (Foster and Kesselman 1999). That is the approach we demonstrate here.

The potential of high performance computing to enable spatially detailed methods of data analysis has long been recognised by those operating across the boundaries of geographical and computational science. Perhaps the most well known work was that undertaken by Stan Openshaw and colleagues, initially at the University of Newcastle and then at the Centre for Computational Geography at Leeds University, especially their pioneering "Geographical Analysis Machine" (GAM) (Openshaw et al. 1987).

GAM worked by passing a moving window across a study region, repeatedly testing for unusual clusters of a particular feature – in the most famous study, clusters of childhood leukaemia, some of which were found in proximity to a nuclear power station. It was an early example of automated and exploratory spatial data analysis made possible by three broad trends. First, the increased availability of high performance "super computers". Second, the proliferation of digital data with point (i.e. x, y) geocoding. Third, the recognition within quantitative geography of the need to move away from statistical techniques that 'smooth over' geographical variation to using more geographically attuned forms of local statistics revealing spatial patterns within data.

Those trends have continued. Computers are ever more powerful. Data have propagated in ways that were hard to imagine 20 years earlier with historically unprecedented access, via digital archives or online portals, to the products of new data gathering and geocoding (such as developments in GPS, remote sensing, surveillance and micromarketing). Spatial statistics have developed and evolved in fields such as geostatistics and spatial econometrics.

What is new is the development of "E-social science", whereby multi processor computer resources are (in principle) available to academic communities for computationally-intensive analyses of large datasets. In the UK, E-social science is led and promoted by the National Centre for E-social Science (NCeSS) and by the National

Grid Service (NGS), the latter of which aims "to provide coherent electronic access for UK researchers to all computational- and data-based resources and facilities required to carry out their research, independent of resource or researcher location" (see http://www.grid-support.ac.uk for additional details). In other words, they provide access to distributed computing.

E-social science is a fledgling and multidisciplinary research field that offers new opportunities for geographical and statistical research. Writing about its prospects for geocomputation, Martin (2005) identified four essential research issues for E-social science. They were: (1) automated data mining; (2) visualization of spatial data uncertainty; (3) incorporation of an explicitly spatial dimension into simulation modelling; and (4) neighbourhood classification from multi-source distributed datasets.

Missing from Martin's list but incorporating elements of data mining, spatial data uncertainty, and spatial data modelling is the potential of E-social science to lead and bring together a whole toolkit of spatial statistical methods in a common but distributed computing environment, within which the analysis of large geographical datasets may be undertaken at ease. This article is a first step in realising such potential, reporting on how one method of spatial analysis, Geographically Weighed Regression (GWR), was implemented within the distributed computing infrastructure provided by the NGS, and reflecting on the limitations of the "grid-enabled" approach.

In the article, we present the rationale for grid-enabling GWR, present a case study of its use looking at participation in Higher Education in England, identify some unresolved issues, and consider why problems in future development might warrant a change of direction. We begin with a discussion of GWR, exploring why, as a method of local spatial statistical analysis, its implementation as a grid service provides a benchmark for other methods of spatial analysis.

2 GWR as a Method of Local Spatial Statistical Analysis

The theory of Geographically Weighted Regression (GWR) and its foundations in more conventional methods of regression analysis are developed in detail by Fotheringham et al. (2002). They are well summarised by Nakaya (2008).

Here it is sufficient to note that GWR is a method of exploratory data analysis that allows the user to reveal geographical variations in the relationships between a dependent variable and one or more predictor variables. Whereas traditional methods of regression analysis are essentially non-geographical, assuming, for example, that the model's residuals are independent of each other and that the modelled relationships are stationary across geographical space, GWR begins with the opposite view, anticipating spatial dependency but also spatial variation. It is a model of spatial heterogeneity.

In particular, GWR begins with the expectation that the predictor variables will vary continuously and spatially in respect to their effects upon the dependent variable. The relationships therefore are assumed to be non-stationary. In this way, a regression equation of the form:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad [1]$$

becomes

$$\hat{y}_{(u_i,v_i)} = \beta_{0(u_i,v_i)} + \beta_{1(u_i,v_i)}x_1 + \beta_{2(u_i,v_i)}x_2 + \ldots + \beta_{k(u_i,v_i)}x_k \quad [2]$$

with the coefficients for each of the predictor variables now assumed to vary from point i to the next across the two-dimensional geographical space defined by the grid coordinates (u, v) (or coordinates on a sphere).

The change of perspective means GWR can be used to explore and detect geographically localised deviations from more broadly expected, that is, "global" trends. GWR also provides statistical explanation about why those local variations exist and where they are statistically significant. In short, GWR generates empirical evidence for where spatial patterning exists – evidence that may validate prior expectations and theory, or raise further questions about the processes and structures operating within the researcher's chosen study region (especially where they invalidate the assumption of residual independence underpinning traditional regression).

Most importantly for this article, GWR is emblematic of other methods of local spatial statistical analysis in at least six regards.

First, its use can be exploratory, used to detect geographical variations and differences. The underlying epistemology is that "geography matters" (or, at least, that it might, and that it might should be examined).

Second, and following from the first, GWR is used for data that are geocoded with point coordinates in a two-dimensional geographic space. Based on those coordinates, a search window (a kernel) is centred on, and passes sequentially from one (fit) point to the next across the study region.[1] As it does so, local statistics are generated. For GWR these are the results of fitting distance weighted regression models. Data points that are outside the search window are weighted as zero. In effect, then, the search window defines a series of spatial subsets of the data.

Third, the statistical analysis around each fit point is undertaken independently of the others. The result for one location does not affect the result for another: they are separate analyses. Consequently, the order of analysing the fit points is irrelevant. In every case, the underlying algorithm or sequence of commands used to analyse the data remains the same. In this way, GWR can be likened to a constant function into which only the weights attached to the data change from one function call to the next. The order in which the sets of data are passed to the function does not matter.

However, and fourthly, the way the search window is defined *does* matter. Spatial statistics characteristically place a search window, or kernel, over contiguous sub-regions of the study area, fixing the size of the search window (the radius of the kernel) or holding constant the number of observations in each subset (which accommodates variations in the underlying population density). The size of the search window or kernel is often described as the "bandwidth".

Larger bandwidths contain more data. The size of the bandwidth must therefore affect the results. This, classically, is the modifiable areal unit problem (MAUP), which includes the issue of scale. If the bandwidth covers too great an area then it risks missing or "smoothing out" important local variations in the attribute data and in their relationships; too small, and the potential for errant data to adversely affect the results is high. It is a balance between precision and accuracy (also, standard error and bias). As a consequence, either the bandwidth should be defined *a priori* with a value that has some theoretical justification or, more usually and consistent with the exploratory nature of the analysis, a process of calibration is required, using a cross-validation or other optimisation procedure to determine the bandwidth's spatial extent.

Fifth, having analysed the fit points separately, the results are pooled and compared. For types of hot spot" analysis this could be a comparison of the local means and standard deviations against each other or against the global values for the entire study region (each a form of t-test). For GWR, the regression coefficients calculated at each of the point locations are examined for evidence of spatial variation. A simple way to do this is to summarise the distribution of the coefficients using quantiles and to identify those parameters of the model with the greatest interquartile range. This is the procedure incorporated in the desktop GWR software available from the National Centre for Geocomputation, National University of Ireland, Maynooth. Later in this article we will also use a form of quantile comparison but with deciles, reflecting the greater amount of data and not wishing to cut too much from the tails of the distribution.

Sixth, GWR does not scale well, taking considerable time to complete for (not especially) large data sets. Unfortunately, this delay conflicts with the exploratory nature of the analysis and the idea that the user is able, in some sense, to "interact" with the data, learning things of it and drawing out interesting patterns and trends. For example, based on trials using a data set of n = 100,000 point locations (and five predictor variables) we estimate it would take two weeks or more to complete the analysis using an open source implementation of GWR (the spgwr package, see below). To the best of our knowledge, the largest dataset for which GWR previously has been published was of size n = 12,493 (this being the house price study of London reported in Fotheringham et al. (2002)).

Yet, despite the similarities – summarised in Table 1 – there is a key difference between GWR and many other forms of local spatial statistical analysis that adds to the scaling problem: it fits weighted regression surfaces, many of them! This process of repeated regression – and, specifically, the fact that the bandwidth must be optimised – is the main reason why GWR is so computationally demanding and time consuming, and more so than most other techniques. It is also the reason why GWR provides a benchmark for other forms of spatial statistical analysis. If GWR can be enabled to run in reasonable time on large data sets within a grid environment, then computationally simpler methods will run faster still. Usefully, GWR offers a proof of concept for other spatial statistical analyses.

3 Grid-enabling GWR

In the preceding section we identified the problem of using GWR with large data sets: it does not scale well to a large n. Here we outline a solution, taking advantage of the fact that one set of weights and the data can be analysed independently of the others. The independence allows for simple parallelisation of the method, for "grid-enabling" GWR.

To elaborate, consider a dataset of length n, where n is the number of rows and is equal to about 10^5. Assume that each of the rows represents a point location and that the locally weighted regression coefficients will be estimated at each of those locations. That implies 10^5 distance weighted regression surfaces need to be fitted and, because of the distance weighting, an $n \times n$ distance matrix be calculated (giving the distances between each pair of points).

To reduce the computational burden, a heuristic could be sought, taking advantage of the fact that points located a long way apart are unlikely ever to be found in the same

Table 1 Six similarities between GWR and other methods of local spatial statistical analysis, suggesting why "grid-enabling" GWR provides a benchmark for other (simpler) methods

	Characteristic	Explanation
1	Exploratory	Used to detect geographical variations and differences
2	Operates using point data	Observations are analysed based on their location in a two-dimensional grid space
3	A form of local analysis. The data around any one location can be analysed independently of other locations	The method is characterised by repeat testing – sets of weighted data are created and analysed in sequence. The order in which the sets are analysed does not matter.
4	Affected by the MAUP	The bandwidth defines the number of subsets and the data they contain, and so directly affects the end results. Some calibration or optimisation of the bandwidth usually is desirable.
5	Comparative	The results obtained around each point are pooled and compared to look for variations across the study region.
6	Scaling issues	The time taken to complete the analysis rises exponentially with the length (n) of the dataset. This especially is a problem for GWR because of the repeat fitting of regression surfaces.

search window and therefore the distance between them is not required. Whilst potentially true, this is difficult to confirm in advance of the bandwidth calibration and, in any case, offers only marginal savings in terms of the overall computational load. Unfortunately, the main bottleneck for GWR is not the derivation of the distance matrix but the regression fittings. The matrix decompositions involved are of the order n^3 to calculate, whereas the distance matrix is of the order n^2.

Worse, unless the bandwidth can be specified *a priori* then there are actually more than n regression surfaces to fit. The actual number is $n \times m$, where m is the number of iterations required to optimise the bandwidth and is of unknown value until the optimisation is completed. If m is found to be in the range from 10^1 to 10^2 (which is usual) then there could be as many as 10^7 regression surfaces to fit, requiring in the order of 10^{21} calculations to complete.

The solution is to distribute those calculations across multiple processors. To understand how to do this, begin by considering an implementation of GWR on a single processor machine. Assuming the distance matrix has been calculated, a logical way of proceeding with the analysis is to take an initial bandwidth, fit the regression surface at

location 1, then move on to location 2, then 3, 4 and so on until the nth location. At this stage, the bandwidth would be changed, and the entire sequence repeated m times, until the bandwidth has been optimised.

Recall, however, that the locations could be sequenced in any order because the results are independent. It follows that if there are two processors available then the data can be analysed two locations at a time (and in any order). If there are three processors available, then three locations can be considered concurrently; the logic extends in the same way to the total number of processors available for the analysis (a number that we shall denote as k).

In this regard, GWR is an "embarrassingly parallel" process. Parallel, because the data can be sent to separate processors and analysed simultaneously. Embarrassing, meaning simple, because the processes do not interact: each set of data is analysed independently of the others, only afterwards are the results pooled and compared. In broad terms, the time, t, taken to complete the GWR analysis will be inversely proportional to k. Therefore, a doubling of k leads to a halving of t. The way to scale GWR to large datasets is to have multiple processors available to work on the analysis. This precisely is what a grid infrastructure offers.

Furthermore, the use of parallelisation applies not only to the optimisation of the bandwidth but also to the derivation of the distance matrix. This is because an $n \times n$ matrix can be split into n rows (vectors), each of size $1 \times n$. Consequently, a first processor can calculate the distance from location 1 to all the other points whilst, simultaneously, a second processor is calculating the distance from location 2 to all others, and so forth. Once each row has been calculated separately the complete $n \times n$ finally can be assembled. Again, the time required to calculate the matrix will be inversely proportional to k. We ignore the possibility of further halving t on the basis that the distance from 1 to 2 is the same as from 2 to 1 (etc.) because we wish to avoid communication between the processes or storing too much information within a memory (see below).

In fact, the two processes of calculating the distance matrix and of fitting a distance weighted regression model can be combined as a single function operating independently on each of the processors available. The first processor receives a copy of the data, calculates the distances from location 1 to all others and then fits the regression model on the basis of some starting bandwidth. The second processor also receives a copy of the data, calculates the distances from location 2 and fits the regression model using the same bandwidth. The third processor does the same for the third location, and so forth. When all the locations have been analysed, the results are pooled and an assessment made to see if an optimal bandwidth has been achieved. If not, then the function call is repeated but with a different bandwidth value.

Reading the above again carefully will reveal an apparent inefficiency. Observe that the function calculates both the distance matrix and the regression surfaces *each time it is called*. Changing the bandwidth requires the regression surfaces to be refitted (because their distance weighting has changed) but why recalculate the distance matrix when the distances between the points have not changed? And why do so a total of m times?

The answer concerns the definition of efficiency. An algorithm for use in a highly distributed computing environment needs to consider not just processor availability but also memory, persistence of data, and bandwidth. Consider, first, the resource required to store a full distance matrix in memory. For a data set consisting of 100,000 points, the

memory requirement is more than 18 GB (assuming four bytes of memory are required to represent a floating point number). By contrast, if each processor holds only the distance from a specific data point to each of the others, then the memory requirement is 1/k th of the matrix (approximately 180 MB if k = 100).[2]

Next consider data persistence. Although in principle the distance matrix could be calculated just once for the bandwidth optimisation, in practice there is no guarantee within a truly distributed environment that the same hardware will be used for every iteration. This means that unless it is recalculated, the distance matrix will have to be stored by the system controlling the optimisation process. This is possible but not desirable because of potential limitations regarding the bandwidth of the communication paths between the systems within the distributed environment. Even for large datasets, if there are sufficient processors available, then it typically takes less time to recalculate the distance matrices at each iteration than to broadcast each distance matrix to the associated computer node. In general, uncertainty regarding memory and bandwidth restrictions means it is more sensible to discard and recalculate the matrix with each iteration of the bandwidth optimisation than to try and retain it.

4 Implementing Grid-enabled GWR in R

An objective of the NCeSS-funded research upon which this article draws was to create a way of grid-enabling GWR that is as interoperable as possible with existing software. To achieve this, we built on a library for running the GWR computing and statistical package, R (http://cran.r-project.org/). This is the spgwr library, developed by Bivand and Yu (see http://cran.r-project.org/web/packages/spgwr/index.html and Bivand et al. (2008) for additional details), that provides functions for calibrating the bandwidth and for calculating the geographically weighted regression parameters.

A part of the reason for choosing R is that it is open source and freely available. A second was that a previous NCeSS project called SABRE in R had involved the Lancaster University Centre for e-Science developing a parallel implementation of SABRE (a program for the statistical analysis of binary, ordinal and count recurrent events) as R Objects. That project had used middleware called GROWL "to provide user friendly access to GRID resources for applications accessible from desktop computers" (see http://www.ncess.ac.uk/research/quantitative/cqess/growl/ for additional details).

Using GROWL technology, a package was developed that allows GWR to be run on a desktop PC using the existing spgwr library but for which the actual data processing occurs remotely on the National Grid infrastructure. The package was entitled multiR and is actually a client/server system that provides a means of submitting a group of tasks for processing on multiple and remote systems. These systems could be processors on a local high performance cluster, a Condor pool or, in the case of the research, the NGS. The multiR client interface is distributed as a package for R and its usage is similar to that of the standard R function. The idea of multiR is to provide a means of invoking an R function multiple times and with varying arguments, where the result of the function is evaluated on multiple processors. By doing so, R becomes a programming environment for course grained parallel processing.

Figure 1 outlines the principle of multiR and its three tier architecture. Clients use R to define the function and use multiR to submit a job to the multiR server. The multiR

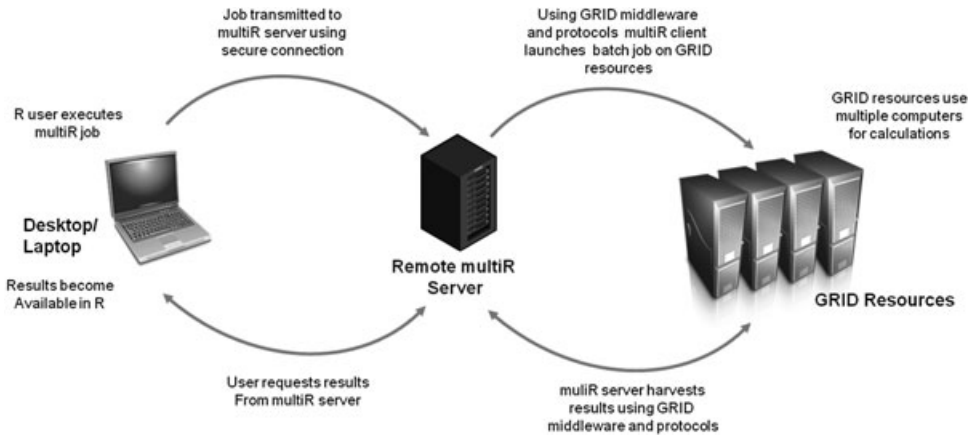

Figure 1 The three tier client/server architecture employed by multiR

server then delegates these tasks to whatever resources it can employ, managing the interface between the user's computer and the NGS.

In addition to multiR, the accompanying spgwr.dist library contains the functions required for grid-enabled GWR (dist is an abbreviation of distributed). A typical session in R begins as:

```
> library(spgwr.dist)
# loads the spgwr.dist and multiR packages
> session <- multiR.session("stats-grid.hpc.lancs.ac.uk",
"50000",
+ "~/mycertificate.p12", "~/multiR.CA.pem ")
# identifies the user's security credentials to use the NGS and
also the multiR server
```

The analysis then continues in a way similar to using the spgwr packgage. Where, in spgwr, the bandwidth for GWR is calculated on the user's desktop using a function of the form:

```
> bandwidth = ggwr.sel(y~x, attribute_data, georeferences)
```

for the grid-enabled version we use:

```
> bandwidth = ggwr.sel.dist(session, y~x, attribute_data,
georeferences, max.processors)
```

Similarly, where the model is fitted in spgwr using:

```
> gwr.model = ggwr(y~x, attribute_data, georeferences,
bandwidth)
```

It is fitted in spgwr.dist using:

```
> gwr.model = ggwr.dist(session, y~x, attribute_data,
georeferences, bandwidth, max.processors)
```

The only difference from the user's perspective is that the additional parameter "session" contains the information required to connect to the multiR server – including the user's security credentials – and the parameter "max.processors" specifies a maximum number of processors the GWR analysis should run on.

5 Modelling Participation in Higher Education in England

In order to demonstrate the use of grid-enabled GWR, a dataset was selected of a size that would be time-consuming to analyse using conventional GWR software. This comprised neighbourhood data at the Lower Super Output Area scale (average total population of 1,500 people), most of which was from the education domain of the 2007 Index of Multiple Deprivation (Noble et al. 2008), see Table 2, below.

The response variable for the analysis is higher education participation rates for the whole of England (defined in Table 2). Interest in this variable arises from the current Labour Government's desire to raise the participation of traditionally underrepresented groups within UK Higher Education Institutes (HEIs), linking to a target for 50% of young adults to be in Higher Education by 2010 (the current rate is about 43%). In turn, the HEIs are required to have access agreements, bursaries and widening participation schemes to support such rises, and government funding is provided to institutions specifically for widening participation initiatives.

Whilst much research has been undertaken looking at socio-spatial inequalities and unequal patterns of access into UK HEIs, (see, for examples, Archer et al. (2003), Reay et al. (2005), and Batey et al. (1999)), still the National Audit Office has reported that:

Table 2 Modelling participation in Higher Education in England: The choice of variables and how they are derived

	Data	Numerator/Denominator	Source
Y	Higher education participation	Successful entrants under 21 in UCAS data, for 2002–2005/Census population aged 14–17 years. The natural log of this value is used in the model to counter a positive skew.	2007 Index of Multiple Deprivation
X_1	No qualifications	Adults aged 25–54 in the area with no qualifications or with qualifications below NVQ Level 2, for 2001/All adults aged 25–54.	2007 Index of Multiple Deprivation
X_2	No post 16 qualifications	Those aged 17 still receiving Child Benefit in 2006/Those aged 15 receiving Child Benefit in 2004.	2007 Index of Multiple Deprivation
X_3	Average KS4 Points	Total score of pupils taking KS4 in 2004 and 2005 in maintained schools from the NPD/All pupils in their final year of compulsory schooling in maintained schools for 2004 and 2005 from PLASC.	2007 Index of Multiple Deprivation
X_4	Four or more cars	Four or more cars in household/total households	2001 Census
X_5	Asian	Total Indian, Pakistani, Bangladeshi people/total people	2001 Census

> Not enough is known about the extent to which disadvantaged groups are under-represented in higher education, or what measures to widen participation are most effective. The Department and the Funding Council need to secure better data on participation, for example by social class or disability. They could tailor provision more closely to people's circumstances, such as where they live and when they can study.
>
> Comptroller and Auditor General (2008)

In particular, little research has been conducted into the spatial variability of the factors that are known to influence higher education participation rates, perhaps with the exceptions of Batey et al. (1999) and Singleton and Longley (2009). It is an area of study where grid-enabled GWR can usefully be applied.

The explanatory variables in Table 2 are not exhaustive but were chosen to represent those factors identified in the literature as affecting differential higher education participation rates. Three of the variables relate to various aspects of educational attainment within residential neighbourhoods, including the proportion of the population without qualifications, the proportion of the population without post 16 qualifications, and the level of Key Stage 4 attainment (a measure of educational achievement at the end of the period of compulsory schooling). Of these, the proportion of the population without qualifications reflects how those people living in areas of low educational attainment are less likely to have been supported throughout their educational careers and also less likely to have associations with others with experience of higher education. Burnhill et al. (1990) illustrate this as an important link to non-participation in higher education by demonstrating a strong association between parental education, specifically to higher education level, and the probability that their children attend higher education.

The variables no post 16 qualifications and average Key Stage 4 points represent the underlying potential for areas to supply qualified candidates eligible for higher education. Possessing a post 16 qualification is usually a minimum requirement for entry to most courses of higher education and low attainment in general has been found to be a further factor affecting participation in higher education across numerous previous studies (Vernon et al. 2002, Gillchrist et al. 2003). Participation in higher education is more likely to occur in higher income families for a variety of reasons such as attitude to debt (Callender and Jackson 2005). However, income information is not collected on the decennial censuses, and as such, surrogate measures are often used. One such substitute for income is car ownership (Dargay 2001); however, it is expected that this variable will demonstrate geographical instability, most specifically in metropolitan areas where alternate transportation is more appropriate (Longley and Toban 2004).

The final variable included in the model is related to ethnicity, known to be associated with differential rates of higher education participation (Modood and Acland 1998, Reay et al. 2001) and school attainment (Hamnett et al. 2007).

Prior to the geographically weighted regression, a global ordinary least squares regression was fitted of the form of Equation (1) and with n = 31,378 observations. Each is a Lower Level Super Output Area (LLSOA, a Census zone). The results are shown in Table 3. The model has an adjusted R^2 of 0.73 and all the predictor variables are significant at a confidence exceeding 95% (not surprisingly, given the size of n).

We expect the global model to conceal geographical variation in the effects of each predictor variable on entry into Higher Education. Evidence that it will do so is found by producing a Moran plot of the residuals from the regression model with their spatially

Table 3 Results for the global model

	β	Standard error	t value	Significant at a 95% level?
(Intercept)	3.620	0.0213	170.2	Yes
X_1: No Qualifications	–0.027	0.0002	–152.5	Yes
X_2: No Post 16 Qualifications	–0.002	0.0001	–15.1	Yes
X_3: Average KS4 attainment	0.003	0.0002	52.6	Yes
X_4: Four or more cars	0.018	0.0005	35.9	Yes
X_5: Asian	0.012	0.0002	68.1	Yes

lagged equivalents (generated using a first order contiguity matrix). What we find is an overall pattern of positive spatial autocorrelation where a positive residual for one census zone is surrounded by positive residuals for its neighbours, and negative neighbours surround a negative residual, yet also with exceptions to the overall trend. This is shown in Figure 2 for a random sample of the residual data (the sample being used to help limit the over-plotting that occurs with all 31,378 observations).

Therefore, a GWR model was fitted, using the grid enabled system described earlier, a Gaussian weighting scheme and an adaptive kernel. The optimal bandwidth was found to be that containing the 97 nearest neighbours to the geographical centroid at the centre of each LLSOA. The shape of the kernel is therefore Gaussian to the 97th neighbour. Beyond that points will have a weight of zero.

The results of the GWR model are summarised in Table 4. The first column gives the global regression coefficient for each of the predictor variables, repeated from Table 3. The remaining columns indicate how that coefficient varies across the study region. For example, 'on average', a percentage point increase in the percentage of adults without a qualification (X_1) will lead to a 0.030 point reduction on the intake variable. This is the median value in Table 3 and, reassuringly, is similar to the global value. Across the study region, though, the complete set of 31,378 fitted beta values ranges from a minimum of –0.047 to a maximum of –0.014. Focusing on the first and ninth deciles of the distribution, the values range from –0.036 to –0.023. The intequartile range (IQR) is also given; in this case it is equal to 0.006.

Looking at Table 4, the varying effect of the car ownership variable is evident. Comparing the coefficient at the first decile ($\beta = 0.011$) with the coefficient at the ninth decile ($\beta = 0.040$) implies that high car ownership is associated with a rate of participation in higher education that is almost four times greater in some places than others. This variation is driven by a "London effect": 63% of the 3,089 fit points with a coefficient in or above the ninth decile – those with the strongest relationship between high car ownership and participation in higher education – are in Greater London; less than 3% of the 3,025 points with a coefficient at or beneath the first decile are. What this is revealing is the concentration of high wealth within the London region and how that wealth provides direct and indirect pathways into Higher Education.

Still with reference to Table 4, an interesting variable is that indicating an Asian ethnic category: it ranges from being negative to positive in its effect upon the rate of University participation. Figure 3 maps the distribution of those beta values, revealing

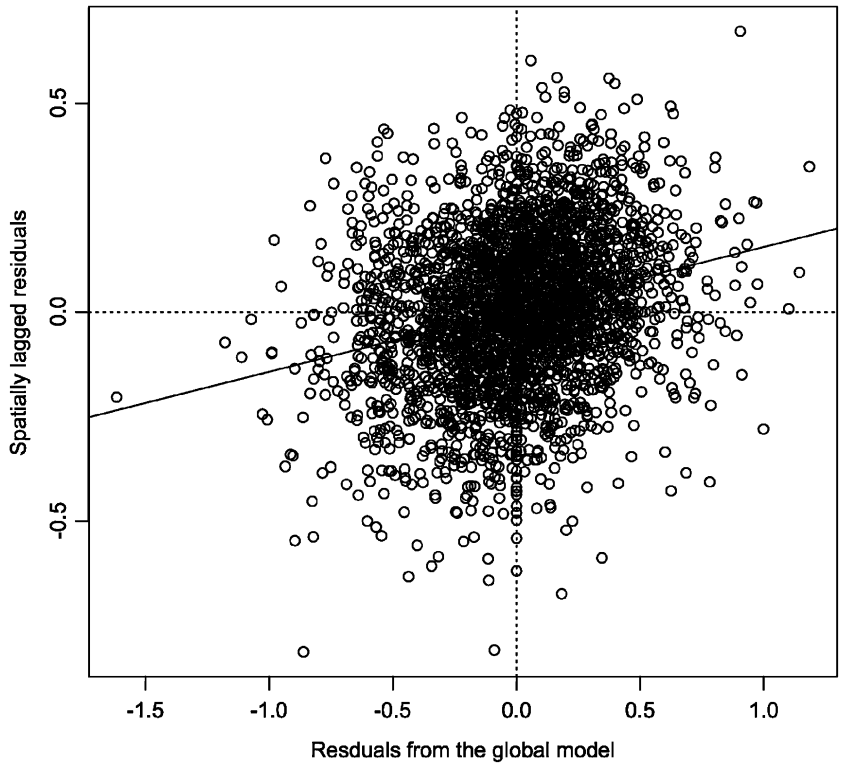

Figure 2 A Moran plot suggesting evidence of positive spatial autocorrelation in the residuals from the global regression model. A random sample of 10% of the data is plotted

some distinct regional and sub-regional differences. However, it would be erroneous to assume that each of these beta values is significant in a statistical sense. Indeed, only the minority (31% of the 31,378 fit points) are significant at a t-value of two or above (none of the negative coefficients is significant at t = –2 or below). Those points where the relationship is significant are shown in Figure 4 and are strongly clustered in the urban industrial conurbations, reflecting the manufacturing geography of the England in the mid 20th century and patterns of immigration.

A notable exception is the cluster found in the South West region of the map, in Cornwall. This cluster is located around Falmouth, an historic but still a cargo port. It is also the location of the Combined Universities in Cornwall where more than £60 million has been invested in providing new academic facilities in one of the most remote parts of the country.

6 Conclusions

In this article we have outlined a technical and reasonably "user friendly" solution that permits GWR to be applied to relatively large data sets, here modelling participation

Table 4 Summary of the results for the GWR model

	β (global value)	$\beta_{(u,v)}$ Min.	$\beta_{(u,v)}$ 1st decile	$\beta_{(u,v)}$ 3rd decile	$\beta_{(u,v)}$ Median	$\beta_{(u,v)}$ 7th decile	$\beta_{(u,v)}$ 9th decile	$\beta_{(u,v)}$ Max.	$\beta_{(u,v)}$ IQR
(Intercept)	3.620								
X_1: No Qualifications	–0.027	–0.047	–0.036	–0.032	–0.030	–0.027	–0.023	–0.014	0.006
X_2: No Post 16 Qualifications	–0.002	–0.008	–0.003	–0.002	–0.001	–0.001	0.000	0.005	0.002
X_3: Average KS4 attainment	0.003	0.000	0.001	0.002	0.003	0.003	0.004	0.006	0.001
X_4: Four or more cars	0.018	–0.013	0.011	0.016	0.021	0.027	0.040	0.101	0.014
X_5: Asian	0.012	–0.156	–0.006	0.009	0.012	0.015	0.020	0.217	0.008

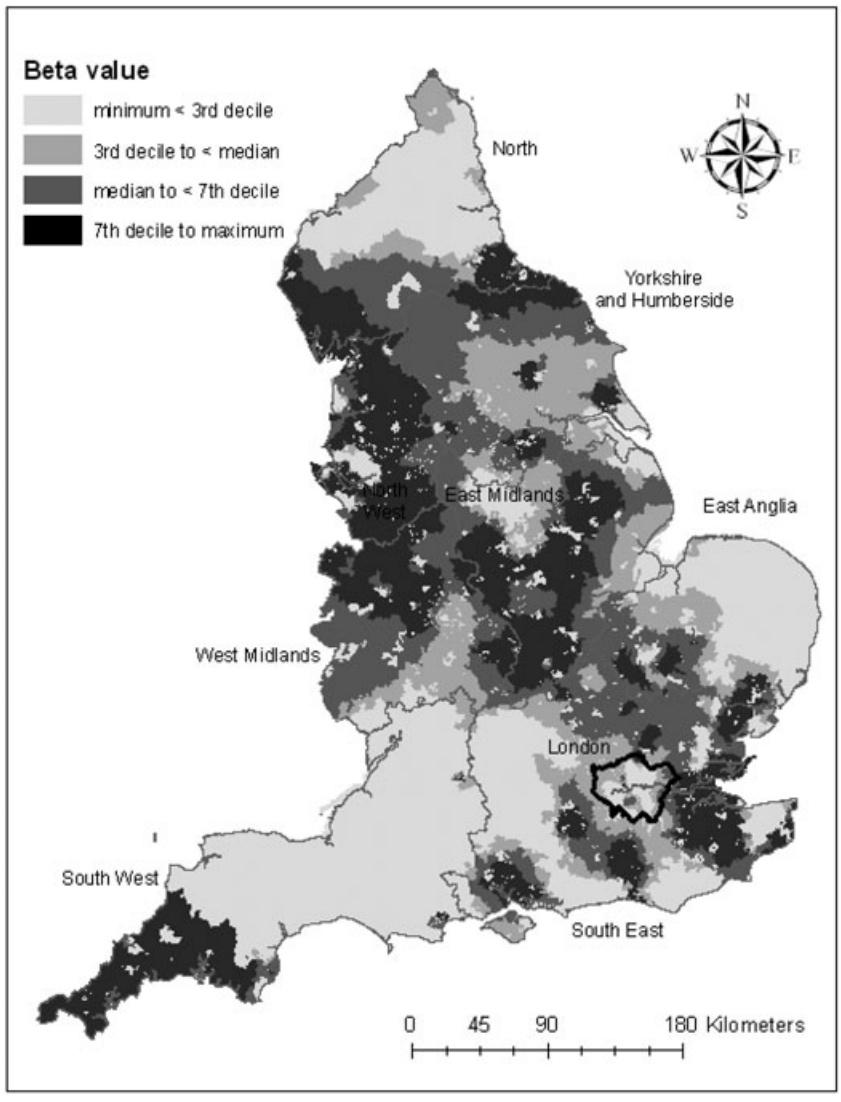

Figure 3 Illustrating the spatial variation in the local (GWR) coefficients for the indicator of Asian ethnicity

rates in Higher Education in England. Unfortunately the solution is not a panacea; we end by describing some of the further issues and problems that are raised by it, including a dependency on what is more centralised than it is distributed computing.

A first issue is practical. The speed-up in applying GWR to large data sets is proportional to the number of processors (nodes) available for use on the grid system. The system used for the analysis here was the North West Grid, based at Lancaster and with over 100 nodes (http://www.nw-grid.ac.uk/). Assuming all were available for our exclusive use, that reduces the time taken to complete the analysis by the order of 10^2 – to an hour or two to complete. That is fast but hardly offers the sort of rapid

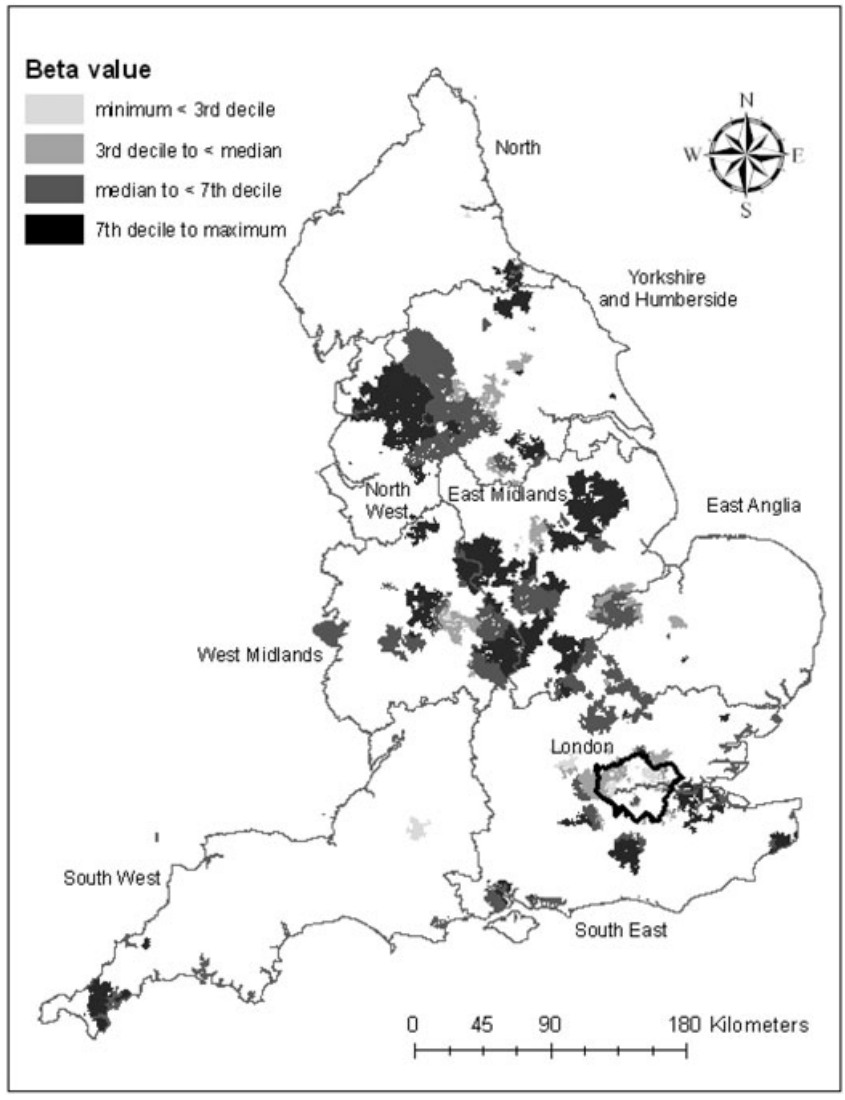

Figure 4 Showing where the indicator of Asian ethnicity is related significantly to the Higher Education participation rate ($t \geq 2$)

interaction with data that exploratory spatial data analysis aspires to. Moreover, it presumes that the user has access to the grid system, which in turn requires the user to be an academic involved in research in the United Kingdom. If they are, they can apply in the first instance via http://www.grid-support.ac.uk/ to obtain a grid user's certificate and to initiate a process of security checking, requiring the production of photo ID to a local administrator, and also a review of the research purpose behind the application. This all takes time, and though it may be necessary for the purpose of secure system administration, it will be off-putting to a more casual but potentially interested user.

The second issue is technical. Figure 3 of this article was based on t-scores, mapping only those fit points where the relationship between the indicator of Asian ethnicity and the rate of participation in Higher Education was deemed significant. Those t-scores were not calculated using grid GWR but using bespoke code to fit local and weighted regression surfaces in desktop R. The reason for this is that R resides in the computer's memory and is limited to about 2 GB for 32-bit Windows. The hat matrix required by R's spgwr library to obtain the local measures of significance more directly will exceed this memory limit for a large data set. As Bivand et al. (2008, p. 10) note, "R may not be suitable for the analysis of massive data sets, because data being analysed is held in memory." That is a general problem that we encounter.

A third issue is how to interpret the GWR model. Increasing the amount of observations input into the analysis raises the prospect of discovering complex geographical patterning. The problem is that all is not held constant as the effect of one predictor on the response variable changes over space. For a model with many independent variables, as one changes in its effect, so the others may be changing too in complex ways. Uses of GWR can be exploratory – examining for geographical differences – and also diagnostic, checking the assumption of spatial stationarity underpinning traditional (OLS) regression analysis is valid. For a more confirmatory approach that seeks to verify or to falsify a theoretical model, it may be better to model departures from a general model and to see how those departures vary spatially, as opposed to fitting many independent but localised regression models and then comparing their coefficients. The former is more the perspective of multilevel modelling and of spatial econometrics, the latter of GWR and of local spatial statistical analysis more generally. Yet, even for the former GWR has a role to play: it may be used to calibrate the general models and the specifications of spatial dependency, of spatial autocorrelation that they employ. Commonly they are based on questionable assumptions of neighbourhood, defined, for example, by first order contiguity. GWR can be used in tandem with these other approaches to provide a better understanding of the spatial structure contained within the data.

Fourthly, and related to the above, grid-enabled GWR raises our aspirations of how to model that spatial structure. Ultimately GWR employs a single specification of spatial autocorrelation – the bandwidth value, whether defined by nearest neighbours or by physical distance. There is no particular reason that this should be the same everywhere and it could be regionalised. That it is not is for practical reasons: to do so requires a greater computational power that only now is available. However, there will always be a trade-off between a generally applicable model that lacks specific detail, and one that appears more attuned to geographical difference but is overly calibrated on one particular data set (and the error and uncertainties that it contains). A challenge of developing new computational tools to better reveal the spatial patterning of the social, economic or natural landscape and to better understand the processes causing that patterning is to know when to stop, to discern when the study has become overly idiographic.

For now, we have demonstrated the successful application of GWR within a grid infrastructure and, having done so, have provided a benchmark for other methods of local spatial statistical analysis. Unfortunately our ending is not a happy one. Despite the talk of distributed computing, the computing infrastructure behind e-science in the UK is, in fact, located in only a few institutions. The idea may be to "plug in" and use their services (for free) but this is far removed from their realities of research funding and the pressure on Universities to generate income streams. Specifically, there is no guarantee the

service will continue to be funded and, in the case of the multiR server, it has been withdrawn, at least for now. Nevertheless, the package remains available at http://e-science.lancs.ac.uk/multiR/ where there is potential for development for use on Condoor pools and other local networks.

7 Acknowledgments

The research presented here was originally funded by the National Centre for E-social science as a small grant project (ESRC RES-149-25-1041). Further and on-going research has been made possible by a SPLINT (Spatial Literacy in Teaching) Fellowship. Further information about multiR can be found at http://e-science.lancs.ac.uk/multiR/.

Notes

1 The fit points need not be the same as the data points. Though they often are, GWR can also be used to interpolate values at locations where data have not been collected.
2 The problem of memory requirement may pass, in time, with technological development.

References

Archer L, Hutchings M, and Ross A 2003 *Higher Education and Social Class*. London, Routledge

Batey P, Brown P J B, and Corver M 1999 Participation in higher education: A geodemographic perspective on the potential for further expansion in student numbers. *Journal of Geographical Systems* 1: 277–303

Bitter C, Mulligan G, and Dall'erba S 2007 Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems*, 9: 7–27

Bivand R S, Pebesma E J, and Gómez-Rubio V 2008 *Applied Spatial Data Analysis with R*. New York, Springer

Brunsdon C, McClatchey J, and Unwin D J 2001 Spatial variations in the average rainfall-altitude relationship in Great Britain: An approach using geographically weighted regression. *International Journal of Climatology* 21: 455–66

Burnhill P, Garner C, and McPherson A 1990 Parental education, social class and entry to higher education 1976–86. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 153: 233–48

Callender C and Jackson J 2005 Does the fear of debt deter students from higher education? *Journal of Social Policy* 34: 509–40

Comptroller and Auditor General 2008 *Widening Participation in Higher Education*. London, National Audit Office

Dargay J M 2001 The effect of income on car ownership: Evidence of asymmetry. *Transportation Research Part A: Policy and Practice* 35: 807–21

Finley A O, Sang H, Banerjee S, and Gelfand A E 2009 Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis* 53: 2873–84

Foster I and Kesselman C 1999 *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco, CA, Morgan Kaufmann

Fotheringham A S, Brunsdon C, and Charlton M 2002 *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, John Wiley and Sons

Fotheringham A S, Charlton M E, and Brunsdon C 2001 Spatial variations in school performance: A local analysis using geographically weighted regression. *Geographical and EnvironmentalModelling* 5: 43–66

Gillchrist R, Philips D, and Ross A 2003 Participation and potential participation in UK higher education. In Archer L, Hutchings M, and Ross A (eds) *Higher Education and Social Class: Issues of Exclusion and Inclusion*. London, Routledge Falmer: 75–95

Hamnett C, Ramsden M, and Butler T 2007 Social background, ethnicity, school composition and educational attainment in East London. *Urban Studies* 44: 1255–80

Huang Y and Leung Y 2002 Analysing regional industrialisation in Jiangsu Province using geographically weighted regression. *Journal of Geographical Systems* 4: 233–49

Longley P A and Toban C 2004 Spatial dependence and heterogeneity in patterns of hardship: An intra-urban analysis. *Annals of the Association of American Geographers* 94: 503–19

Martin D 2005 Socioeconomic geocomputation and e-social science. *Transactions in GIS* 9: 1–3

Modood T and Acland T 1998 *Race and Higher Education*. London, Policy Studies Institute

Nakaya T 2008 Geographically weighted regression. In Kemp K (ed) *Encyclopaedia of Geographic Information Science*. Thousand Oaks, CA, Sage: 179–84

Nakaya T, Fotheringham A S, Brunsdon C, and Charlton M 2005 Geographically weighted poisson regression for disease association mapping. *Statistics in Medicine* 24: 2695–717

Noble M, McLennan D, Wilkinson K, Whitworth A, Barnes H, and Dibben C 2008 *The English Indices of Deprivation 2007*. London, Communities and Local Government

Openshaw S, Charlton M, Wymer C, and Craft A W 1987 A mark I geographical analysis machine for the automated analysis of point datasets. *International Journal of Geographical Information Systems* 1: 335–58

Reay D, David M, and Ball S J 2005 *Degrees of Choice: Social Class, Race, Gender and Higher Education*. Stoke on Trent, Trentham Books

Reay D, Davies J, David M, and Ball S J 2001 Choices of degree or degrees of choice? Class, 'race' and the higher education choice process. *Sociology* 35: 855–74

Singleton A and Longley P A 2009 Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science* 88: 643–66

Vernon G, Damon B, and Davies R 2002 Young people's entry into higher education: Quantifying influential factors. *Oxford Review of Education* 28: 5–20